Journal of Multivariate Analysis 122 (2013) 215–225

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the limiting distribution of the spatial scan statistic

Tonglin Zhang [a,*], Ge Lin [b]

[a] Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2066, United States

[b] Department of Health Services Research and Administration, College of Public Health, 984350 University of Nebraska Medical Center, Omaha, NE 68198-4350, United States

ARTICLE INFO

Article history:
Received 22 February 2012
Available online 19 August 2013

AMS 2000 subject classifications:
primary 62M30
secondary 62J12
62F03

Keywords:
Clusters
Empirical process
Limiting distributions
Kolmogorov–Smirnov test
Spatial scan statistic

ABSTRACT

Bootstrap is the standard method in the spatial scan test. However, because the spatial scan statistic lacks theoretical properties, its development and connection to mainstream statistics has been limited. Using the methods of empirical processes with a few weak regularity conditions, the limiting distribution of the spatial scan statistic, which can provide a theoretical basis for the spatial scan test, is derived. It is shown that the limiting distribution of the spatial scan statistic only depends on the ratio of at risk populations and the collection of cluster candidates, which provides a base to theoretically assess the critical value of the spatial scan test in a real world daily or weekly disease surveillance. A simulation study based on the Kolmogorov–Smirnov test shows that the limiting distribution is consistent with the true distribution. Type I error probabilities and power functions from the limiting distribution and the bootstrap method are almost identical.

1. Introduction

Spatial cluster detection is an important topic in statistics [2,4,6,10]. While there are over a hundred spatial cluster and clustering tests [13], spatial scan and spatial association tests are the most commonly used cluster detection methods. This paper focuses on the spatial scan test because of its wide range of applications in disease surveillance in local government agencies [14,23,24], and its various extensions that are intended to improve computational speed [12], capture irregular cluster shapes [1,17], and account for ecological covariates [28]. However, these new extensions can hardly be compared or integrated, as all of them have been evaluated via the Monte Carlo simulation. Understanding the theoretical properties of the spatial scan test enables us to not only unify different methods of the spatial scan test, but also create a platform for further methodological developments. The resulting distribution will likely improve the efficiency and precision of a spatial scan test at a prescribed condition.

Scan statistics were originally developed for time series data [22]. Suppose the number of events on a given time interval (e.g. [0, 1]) follows a homogeneous Poisson process in the absence of a cluster. A subinterval of a fixed length moves along the time domain so that the number of events contained by the subinterval is maximized. Let $N(t)$ be the number of events observed on $[0, t]$. Then, the scan statistic with the length of $0 < u < 1$ is defined as

$$\sup_{0<t<1-u} [N(t+u) - N(t)].$$

A cluster is detected if the value of the scan statistic is large. Due to multiple testing, the scan statistic is often larger than that expected from intuition when no cluster is present. For this reason, the Monte Carlo method is often used. In order to improve

* Corresponding author.
E-mail addresses: tlzhang@purdue.edu (T. Zhang), glin@unmc.edu (G. Lin).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.08.005

efficiency, such as carrying out the test at a prescribed level, there have been several attempts to derive a precise distribution of the scan statistic for time series data. Examples include the derivation of the exact expression [22], the asymptotic distribution of $N(t)$ under the null hypothesis of uniformity [5], and the approximation [8] of the null distribution.

Although the scan statistic was extended from one- to two-dimensional point data, there has been little work to derive the null distribution with the exception of [3]. Moreover, since most georeferenced disease data are not available at the point level, the most commonly used spatial scan statistic is, therefore, developed at an aggregated level [16]. Suppose a study area has been partitioned into m spatial units and each has an at-risk population and a number of case counts. Suppose C is the only cluster in the study area. Let Y_i be the count, y_i be the observed count, and n_i be the at-risk population in unit i, for $i = 1, \ldots, m$. Assume Y_i satisfies

$$Y_i \sim \text{Poisson}(\theta_i n_i), \quad i = 1, \ldots, m, \tag{1}$$

where θ_i are unknown disease rates. Then, the null hypothesis is specified as

$$H_0 : \theta_i = \theta_0,$$

and the alternative hypothesis is specified as

$$H_1 : E(Y_i) = \theta_{0C} n_i (1 + \delta_i),$$

with $\delta_i \neq 0$ if $i \in C$ and $\delta_i = 0$ if $i \in \bar{C}$ [9,16]. Here, θ_{0C} is the average disease rate for units outside of the cluster. If $\theta_i > \theta_{0C}$, then unit i is within a hot spot; if $\theta_i < \theta_{0C}$, then unit i is within a cool spot. If H_0 holds, then $\theta_{0C} = \theta_0$ is the average disease rate for all units, the same one for each unit.

Let $\mathcal{C}$ be the collection of cluster candidates. For a selected $C \in \mathcal{C}$, let $Y = \sum_{i=1}^m Y_i$, $y = \sum_{i=1}^m y_i$, $n = \sum_{i=1}^m n_i$, $Y_C = \sum_{i\in C} Y_i$, $y_C = \sum_{i\in C} y_i$, $n_C = \sum_{i\in C} n_i$, $Y_{\bar{C}} = \sum_{i\in \bar{C}} Y_i$, $y_{\bar{C}} = \sum_{i\in \bar{C}} y_i$, and $n_{\bar{C}} = \sum_{i\in \bar{C}} n_i$. Then, y, y_C and $y_{\bar{C}}$ are the observed values of Y, Y_C and $Y_{\bar{C}}$, respectively. Assume $\theta_i = \theta_C$ if $i \in C$ and $\theta_i = \theta_0$ if $i \in \bar{C}$. Under $H_0 : \theta_C = \theta_0$,

$$Y_i \sim \text{Poisson}(\theta_0 n_i), \quad i = 1, \ldots, m. \tag{2}$$

Under the alternative hypothesis of hot spot only,

$$Y_i \sim \text{Poisson}(\theta_0 n_i), \quad i \in \bar{C}; \quad \text{or} \quad Y_i \sim \text{Poisson}(\theta_C n_i), \quad i \in C,\ \theta_C > \theta_0. \tag{3}$$

Then, the likelihood function is

$$L_C(\theta_0, \theta_C) = \left(\prod_{i=1}^m \frac{n_i^{Y_i}}{Y_i!}\right)\left(\prod_{i\in C} \theta_C^{Y_i} e^{-\theta_C n_i}\right)\left(\prod_{i\in \bar{C}} \theta_0^{Y_i} e^{-\theta_0 n_i}\right). \tag{4}$$

The likelihood ratio statistic is

$$\Lambda_C = \frac{\max\limits_{\theta_C > \theta_0} L_C(\theta_0, \theta_C)}{\max\limits_{\theta_C = \theta_0} L_C(\theta_0, \theta_C)} = \left(\frac{Y_C/n_C}{Y/n}\right)^{Y_C}\left(\frac{Y_{\bar{C}}/n_{\bar{C}}}{Y/n}\right)^{Y_{\bar{C}}}, \tag{5}$$

when $Y_C/n_C \geq Y_{\bar{C}}/n_{\bar{C}}$ and $\Lambda_C = 1$ otherwise. Since $C \in \mathcal{C}$ is unknown, the spatial scan statistic is defined by

$$\Lambda = \max_{C\in\mathcal{C}} \Lambda_C. \tag{6}$$

In contrast to the scan statistic for time series data, the spatial scan statistic requires the consideration of additional issues. First, a cluster for time series data is a connected subset which must be a subinterval. A spatial cluster may have many shapes, such as circular, elliptical, and irregular polygon, and they cannot be simultaneously considered in the definition. Second, the derivation of the scan statistic for time series data relies only on the length and the center of subintervals, but these two conditions are not sufficient in the derivation of a spatial scan statistic. Third, existing methods for the derivation of theoretical properties of the scan statistic for time series data rely on the construction of a stochastic process [19], which is not directly applicable to the spatial scan statistic. For this reason, the bootstrap method has long been used to compute the p-value of the test statistic.

Since the number of units is finite, the number of possible spatial candidates $\mathcal{C}$ is also finite being smaller that, both of the same magnitude order as, the number $2^m - 1$ of possible non-empty subsets out of m spatial units. This makes the computation of Λ difficult, especially when a bootstrap method is used. In this paper, we develop an approach to the limiting distribution of the spatial scan statistic, which can avoid the use of the bootstrap method in the derivation of the p-value and reduce the computational burden of the spatial scan test.

We used the method of empirical processes to derive the limiting distribution of the spatial scan statistic. It assumes that expected counts are not close to zero in most spatial units and the conditional distribution of the likelihood ratio statistic is approximately χ^2 distributed. By sweeping the likelihood ratio statistic over cluster candidates, the asymptotic distribution of the test statistic is derived. The proof uses the empirical processes method [27, p. 260]. We show that the spatial scan statistic converges in distribution to the square of the supremum of the absolute value (or positive part) of a Gaussian random field under some regularity conditions. This value can be used to compute the p-value of the test statistic.